Vrije Universiteit Amsterdam	Statistical Data Analysis, Resit Exam
Faculty of Sciences	2 July 2021

Use of a basic calculator is allowed. Programable calculators, and communication devices such as mobile phones, smart watches, etc., are not allowed. This exam consists of 5 questions on 4 pages (45 points).

Please write all answers in English. Grade = $\frac{total+5}{5}$**.**

You have 165 minutes (2h and 45min) to write the exam.

GOOD LUCK!

Question 1 [7.5 points]
Indicate for each of the following statements whether it is correct or incorrect. ***In either case****, briefly motivate your answer.*

a. [1.5 points] The Epanechnikov kernel is generally a good choice for kernel density estimation.

b. [1.5 points] The bootstrap does not always provide a good approximation for the distribution of an estimator, an exception being the sample mean, whose distribution can always be well approximated via the bootstrap.

c. [1.5 points] A random permutation procedure (when applicable) can be used the control of the type I error probability of hypothesis tests.

d. [1.5 points] The χ-square goodness-of-fit test can be used to test if a sample comes from a χ-square distribution with a pre-specified number of degrees of freedom.

e. [1.5 points] A leverage point is always an influence point.

Question 2 [7 points]
Consider a random sample $X_1, \ldots, X_n$ from a distribution with probability density function (PDF) f. Let ϕ and Φ respectively represent the PDF and the cumulative distribution function (CDF) of a standard Normal random variable. Consider the following kernel density estimator for f:

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{t - X_i}{h}\right).$$

a. [2 points] Suppose that you are actually interested in estimating the CDF F of the distribution of the data, and so you integrate $\hat{f}(t)$ to get the estimator

$$\hat{F}(t) = \int_{-\infty}^{t} \hat{f}(s)\, ds = \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{t - X_i}{h}\right)$$

Consider also the empirical CDF, denoted here by

$$\tilde{F}(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq t\}.$$

Explain: i) why would you prefer $\tilde{F}(t)$ over $\hat{F}(t)$, **and** ii) why would you prefer $\hat{F}(t)$ over $\tilde{F}(t)$.

b. [2 points] Explain the relation (if any) between the observations X_i having a normal distribution, and the use of the normal kernel ϕ.

c. [3 points] Describe step-by-step how to obtain a bootstrap-based, two-sided $(1-\alpha)100\%$ confidence interval for $f(t)$, where $\alpha \in (0,1)$ and $t \in \mathbb{R}$.

Question 3 [11 points]
A food delivery service is advertising that they manage to deliver food to the majority of their customers in Amsterdam within 20 minutes of the order being placed. In your experience you always wait longer than this, so you decide to enlist the help of a few colleagues to test out the service's claim. You ask 12 colleagues to check the company's website to see how long their last order took and they get back to you with the following: 19.12, 19.13, 20.03, 20.23, 20.38, 21.23, 21.91, 22.71, 25.07, 25.27, 32.00, 35.05 (time is in minutes and the sample has been sorted.)
You interpret the claim being made as the company stating that the median delivery time (call it m) does not exceed 20 minutes so that you want to test $H_0 : m \leq 20$ versus $H_1 : m > 20$. *For the purpose of answering this question you may assume that the sample above is a random sample and that the true underlying distribution of the observations has a unique median.*

a. [4 points] State: i) the test statistic that you would use to perform a sign test to test the hypotheses stated above, and ii) the distribution of this test statistic under the assumption that $m = 20$.

b. [4 points] i) Perform the sign test at significance level $\alpha = 5\%$ using Table 1 by computing the p-value of the test, and ii) explicitly state your conclusion in terms of the issue at hand.

c. [2 points] The Wilcoxon signed rank test is typically more powerful but requires a further assumption. Answer the following: i) what is this extra assumption? and also ii) explain, briefly, how you would investigate whether this assumption is satisfied or not.

d. [1 point] Both the sign test and the Wilcoxon signed rank tests are distribution free. Explain, briefly, what that means.

	p						
k	0.025	0.05	0.33	0.5	0.67	0.95	0.975
0	0.738	0.540	0.008	0.000	0.000	0.000	0.000
1	0.965	0.882	0.057	0.003	0.000	0.000	0.000
2	0.997	0.980	0.188	0.019	0.000	0.000	0.000
3	1.000	0.998	0.403	0.073	0.004	0.000	0.000
4	1.000	1.000	0.641	0.194	0.018	0.000	0.000
5	1.000	1.000	0.829	0.387	0.063	0.000	0.000
6	1.000	1.000	0.937	0.613	0.171	0.000	0.000
7	1.000	1.000	0.982	0.806	0.359	0.000	0.000
8	1.000	1.000	0.996	0.927	0.597	0.002	0.000
9	1.000	1.000	1.000	0.981	0.812	0.020	0.003
10	1.000	1.000	1.000	0.997	0.943	0.118	0.035
11	1.000	1.000	1.000	1.000	0.992	0.460	0.262

Table 1: Probabilities $P(X \leq k)$ for binomially distributed random variable X with parameters $n = 12$ and p as given in table, for different values of k.

Question 4 [9,5 points]
Suppose that you regularly play a board game with friends but for about a year now you have been playing the same game but online. Table 2 below summarises the number of times that you have won/lost in the last 23 games that you played since you started keeping track of the results.
You noticed that you seem to be winning comparatively less often since you started playing online and you are curious if playing online is having a demonstrable impact on your performance. To figure this out you will carry out a statistical analysis of your performance.
Hint: You may need to make use of Tables 4 and 5 to answer some of the questions below.

Outcome	Lose	Win	total
Online	4	5	9
In person	4	10	14
total	8	15	23

Table 2: Your wins and losses for 23 games.

a. [1.5 points] Of the models for multinomial distributions that we discussed (A, B, and C), which is the appropriate one for this problem? State the corresponding null and alternative hypotheses that are being tested. (You may do this either in words or in formulas.)

b. [4 points] i) Test the null hypothesis of part a. at significance level $\alpha = 5\%$ using the fact that the value of the test statistic of the chi-square test is $\chi^2 \approx 0.11$. Do not forget: ii) to check the rule of thumb but perform the test in any case, and iii) to formulate the conclusion of the testing procedure in terms of the question at hand.

c. [2 points] Table 3 contains the standardised residuals associated with the contingency table above. Do the following: i) briefly explain how the residuals are obtained, and ii) indicate which (if any) of the cells can be considered extreme (at significance level 5%.)

Outcome	Lose	Win
Online	0.78	-0.78
In person	-0.78	0.78

Table 3: Standardised residuals for the χ^2 statistic.

d. [2 points] In class you also learned about Fisher's exact test. Answer the following: i) Why is Fisher's exact test applicable in the context of the present question? ii) What is the value of the test statistic for Fisher's exact test for the contingency Table 2 above? iii) Which of the two tests (chi-square test or Fisher's exact test) is preferred in the context of this specific problem and why?

	α						
k	0.025	0.050	0.330	0.500	0.670	0.950	0.975
1	0.00	0.00	0.18	0.45	0.95	3.84	5.02
2	0.05	0.10	0.80	1.39	2.22	5.99	7.38
3	0.22	0.35	1.55	2.37	3.43	7.81	9.35
4	0.48	0.71	2.36	3.36	4.61	9.49	11.14

Table 4: α-quantiles of χ^2_k-distribution for indicated values of α and k.

α						
0.025	0.050	0.330	0.500	0.670	0.950	0.975
-1.96	-1.645	-0.44	0	0.44	1.645	1.96

Table 5: α-quantiles of the standard normal distribution for indicated values of α.

Question 5 [10 points]

a. [2 points] Formulate the general multiple linear regression model including its assumptions.

b. [4 points] For each assumption, shortly describe a method to verify that assumption.

c. [2 points] Explain the concept of (multi-)collinearity in multiple linear regression in your own words. Clearly state what collinearity is and why it may cause problems.

d. [2 points] Consider the data shown in Figure 1. The response variable is the number of packages processed by a sorting machine per hour, and the explanatory variable is the weight of incoming packages (in kg). There are 20 observations.

Do you expect any problems when the simple linear regression model (i.e., linear model with one explanatory variable) is fitted to these data? If yes, how would you investigate it/them?

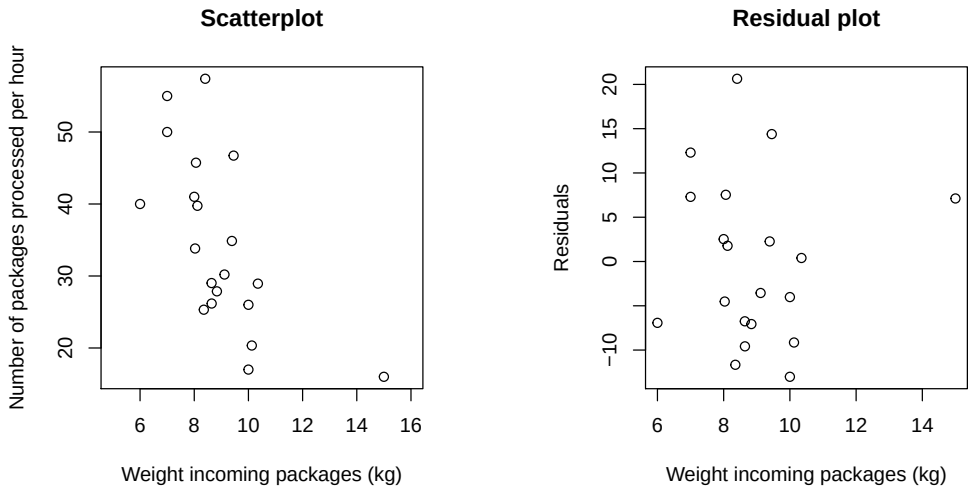

Figure 1: Scatter plot of 'packages per hour' against 'weight of incoming packages' (left), scatter plot of the residuals against 'weight of incoming packages' (right).